Benchmarking Big Data in the Cloud

Key Issues and Challenges

Dan Koren – Actian Corporation

The emergence of cloud hosted BIG DATA as the de facto data storage, management and processing paradigm for web and large scale applications in science and commerce presents new challenges and raises many questions for the performance engineering community –both in academic research and in various computer related industries.

Key issues that need to be investigated and understood include:

New Application Paradigms

Cloud hosted big data typically supports a variety of applications – often cloud hosted as well. Single function one database application servers are a thing of the past. Some HPC and scientific applications seem to be exceptions – which may be regarded as distributed single logical servers. We need to consider both use cases.

Multiple data models of varied types and having different degrees of structure co-exist in the big data clouds: SQL, NoSQL, HTML, documents, multi-media, time series, scientific data. Several application frameworks are common in the cloud – Hadoop/MapReduce, Hive, HBase, Cassandra, SciDB, Amazon's EC2, Actian's own Cloud Action Platform™, and so on. How does one define meaningful workloads that model relevant data models and use cases?

System Scale and Boundaries

What is a SUT (system under test) and how does one determine or define its boundaries? Both the data and the applications that process it are commonly hosted in the cloud. How does one define the boundaries between "clients" and "SUTs"? How can they be enforced? Is the client vs. SUT distinction still meaningful?

How many clients are needed to sufficiently load cloud hosted big data repositories to measure performance in meaningful ways? Unlike classical client-server systems which are usually server limited, processing power in the cloud can easily exceed that of all its clients!

Metrics

What metrics are meaningful? Which metrics would be useful? How should they be normalized and averaged? The number of available dimensions seems overwhelming: nodes, CPUs, memory, connections, disks, users, clients, money, etc…

Which metrics are useful to various benchmark result consumers: CIOs, IT planners, performance analysts, the public at large, industry analysts? Can metrics be defined that are intuitively meaningful and straightforward

to compare across different workloads and systems? The sheer scale of BDIC systems suggests other metrics besides latency and throughput may be highly valuable to users – energy, heat, floor footprints, start or restart time, etc…

Benchmark Definition and Implementation

What workloads should be measured? Should BDIC (big data in the cloud) benchmarks measure single or multiple, mixed workloads? Should one benchmark an application? A system? A framework? An application?

Which layers of the stack can or should be legislated? Can the legislation be kept straightforward, readable and clear to interpret? Should there be a single implementation of each benchmark, à la SPEC? Or should multiple implementations be allowed, à la TPC?

How does one define and enforce calibration, compliance, repeatability and comparability? What about the cost of benchmarking? Can BDIC benchmarks be audited in manageable amounts of time and with realistic budgets?

Measurement and Tools

Accurate time measurement forms the cornerstone of any performance benchmark. While straightforward conceptually, it is not trivial to implement even in simple client-server systems. Measuring time consistently across tens or hundreds of nodes in a cloud raises the bar tremendously and requires sophisticated software and measurement techniques. My talk will focus on methods for accurate time and energy measurement in large scale distributed systems, and their application to benchmarking BIG DATA IN THE CLOUD.

Standards and Industry Forums

How will BDIC benchmarks be defined? Should this be attempted within the framework of existing standards forums, such as TPC or SPEC, or should a new industry consortium be formed?

So many questions to ponder…. So few ideas…. So little time….

Actian Corporation is an industry leader enabling corporations to take immediate action on big data through its revolutionary Action Apps™, Cloud Action Platform™ and Vectorwise Analytical Database™. Dan Koren has served the computer system and database industries far longer than he would like to admit, making every product he touched run a lot faster. He is responsible for Actian's Performance Engineering Program. Dan is a certified trouble maker and rabble rouser who can easily talk his way into any situation. He enjoys acting as a catalyst, devil's advocate and enabler for new ideas. Even his opinions have opinions!

Dan will attempt to answer some of the above questions in a short but provocative slide presentation that is likely to surprise the audience and raise his blood pressure – as well as everyone else's! ☺